Proceedings of the 2002 IEEE
Int. Workshop on Robot and Human Interactive Communication
Berlin, Germany, Sept. 25-27, 2002

CORA: An Anthropomorphic Robot Assistant for Human Environment

Ioannis Iossifidis[1], Carsten Bruckhoff, Christoph Theis,
Claudia Grote, Christian Faubel, Gregor Schöner
Institut für Neuroinformatik, Ruhr-Univerität Bochum, Germany
[1]Phone +49 234 3225567, Email: Ioannis.Iossifidis@neuroinformatik.ruhr-uni-bochum.de

Abstract

We describe the general concept, system architecture, hardware, and the behavioral abilities of CORA *(***Co***operative* **R***obot* **A***ssistant, see Fig. 1), an autonomous non mobile robot assistant. Outgoing from our basic assumption that the behavior to perform determines the internal and external structure of the behaving system, we have designed* CORA *anthropomorphic to allow for human-like behavioral strategies in solving complex tasks. Although* CORA *was built as a prototype of a service robot system to assist a human partner in industrial assembly tasks, we will show that* CORA*'s behavioral abilities are also conferrable in a household environment. After the description of the hardware platform and the basic concepts of our approach, we present some experimental results by means of an assembly task.*

1 Why anthropomorphic autonomous service robots?

Robots - in the sense of human-like general purpose machines that are able to perform well in a human environment and can handle the various badly defined everyday tasks we humans spend a lot of time with - are still a matter of the future.

Besides the fascination to create our own technical analogon, the design of such robots poses a number of problems that are crucial for a lot of applications. The methods that need to be developed as a presupposition for successful robot design are important for many fields that at first glance have nothing to do with robotics. This includes automatic data acquisition, process control, user assistance for handling complex tasks like e.g. driving, and certainly entertainment. Robot design calls for a very high integration of mechanical, electrical, electronical and computational technology.

Contrary to how most robots are constructed today, a look at the biological prototype reveals that not the extensive use of highly precise sensors is required, but that a few less perfect but

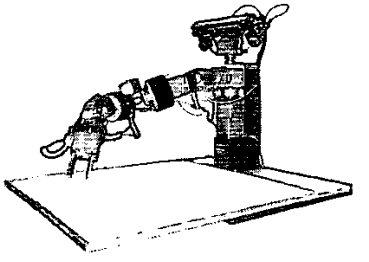

Fig. 1. The service- and assistance robot CORA. A seven DoF manipulator arm is mounted on a one DoF trunk which is fixed on a table. The robot possesses a two DoF stereo camera head with microphones.

universally applicable input channels - namely vision, audio and touch - can be used to great effect. As useful as a collection of special purpose devices for distance measurement, illumination etc. may be in the short run, they imply additional system load and unnecessary complexity for special cases and thus are a dead end for the development of more generic systems.

Another common misconception for intelligent systems is the call for complete user or operator control. Total and detailed control over every aspect of the system is something that a typical classical computer system offers. This is also the reason why those systems are hard to use, require a lot of knowledge and training and are nevertheless easy to make errors on, i.e. to generate command with a different outcome than that intended by the user. Our interaction with more intelligent "systems" like e.g. a trained dog is very different from that: Although the communication channel - speech and gestures - is highly redundant, the syntax and semantics are mostly ill-defined and the animal is never guaranteed to even obey a given command, the interaction is quite intuitive, robust even under difficult conditions and generally pretty powerful. The dog is easy to deal with because it will generally use its own intelligence and only do what his master requests if the commands are not in contradiction to his own "behavioral goals". A dog that is

commanded *"come here straight"* and runs into a wall because it obeys the command would be called particularly stupid with good reason. The same holds true for a robot. Combining real autonomy of behavior with a user interface may be quite straightforward by defining the command input on the same level as the sensory input and the general behavioral goals; such as e.g. obstacle avoidance.

Besides general purpose service robotics there is also the field of robotics for the disabled. Mobility assistances for the blind and wheelchair-mounted robot arms are only two examples posing a technical challenge for highly robust robot control which allows the user to easily specify tasks. One possible and maybe ideal approach for robotic tools for the disabled would be a library of semiautonomous behaviors that perform tedious subtasks - like e.g. locating an object for grasping or determining trajectories. Thus robotics for the disabled share the concept of behaviors or skills as coarsely parameterized atoms by which more complex tasks can be successfully performed.

Anthropomorphic shape and the fluent, predictable movements of effectors controlled by dynamical systems rather than stiff trajectories also have a strong effect on users which makes working with such a robot anesthetically and emotionally more pleasing.

2 *"Anatomy"* - Body Shape, Sensors and Effectors

As discussed before, one of the main goals followed in the construction of CORA was to design the system as anthropomorphic and adapted to a human environment as possible with currently available and preferably well-established technology.

CORA is fixed on a table and meant to physically interact with a human standing across the table. Its head, which is mounted above the body, has two degrees of freedom (DoF): pan and tilt. The head carries a stereo color camera system and microphones. The vision system performs tasks such as object recognition, gesture recognition and the estimation of the 3D position and orientation of objects.

A second stereo camera head is fixed beside CORA on the table. This second vision system is used to estimate the gaze direction of the human partner. It should be mentioned that the establishing of a second head respeak our philosophy of an anthropomorphic design. But in this case it was necessary to equip the robot with a vision system for which the visibility of the humans eyes is warranted at any time. Therefore for technical reasons we decided to build up a second head.

CORA is developed in the BMBF project MORPHA to demonstrate how service tasks in a human environment can be performed by an autonomous anthropomorphic robot combining loosely coupled dynamical systems realizing basic behaviors to achieve a complex overall system behavior.

2.1 Visual Abilities

An ideal stereo camera head allows pure rotation of all cameras around their main point to avoid parallax errors in determining directions and additionally implements a rotating neckto enable symmetrical configurations for any direction.

In the real world, such a solution would be bulky and expensive, thus we use only two degrees of freedom for pan and tilt , where the fovea camera main points lie at the intersection of the pan and tilt axes. The stereo base is at about 300 mm significantly wider than with human anatomy to compensate for the lower resolution of the fovea cameras with respect to the human fovea.

2.2 Acoustic and Phonetic Abilities

CORA's acoustic system consisting of microphones on the hardware side and the speech recognizer *ears* [7] [8] on the software side forms together with the speech synthesizer called *mbrola* [3] [1] a dialog system.

2.3 Head Configuration

For grasping the relative configuration of body, arm and head is crucial: Humans typically use visually controlled arm and hand movements with the position of the head above and behind the grasping position. This and the fact that most visually controlled grasping is done from the side with the elbow pointing outwards gives an optimal visual control of the situation and especially the most critical part, the grip itself. Therefore CORA's head is mounted above its body and its arm is mounted lateral on its trunk.

2.4 Grasping Space

CORA's grasping space, as a non mobile robot, which body consists of a redundant seven DoF manipulator connected to a one DoF trunk which is fixed on the edge of a table, is restricted to this table. CORA can exploit the redundant eight DoF of the arm trunk configuration that guarantees a high degree of flexibility with respect to manipulation tasks under external constraints. Grasping, for instance, is possible in the whole workspace choosing different arm-postures without the necessity of changing the position or orientation of the end-effector.

The body orientation determined by the trunk DoF is chosen with respect to object to be grasp. The body orientation is aligned perpendicular to the object as long it can be reached by the endeffector. Otherwise the robot turns its body to the objects direction until the endeffector reaches the objects position. In addition to it the robot can also change its configuration from left- to right-handed by the trunk joint.

2.5 Arm Configuration

CORA's arm configuration is equivalent to a broadly simplified model of the human arm with a 3 degree of freedom (DoF) shoulder, an 1 DoF elbow and a 3 DoF wrist. The maximum grasping radius is about 1 m with a maximum load of about 1.5 kg.

The redundancy of the seventh degree of freedom allows the handling of situations in which additional movement constraints have to be met. Our main interest in utilizing this redundancy is in obstacle avoidance during grasping, which is important for handling of objects on tables, cupboards, etc.

This human-like arm configuration also has the advantage to be controlled by a distributed control scheme: the wrist position (3 DoF), the wrist orientation (3 DoF), and the elbow angle with respect to the shoulder-wrist axis (1 DoF) can be controlled by 3 independent control loops, allowing e.g. the elbow angle control to use a very specific and efficient obstacle avoidance criterion.

2.6 Haptic Interface

Two of CORA's arm-modules are covered with the so called *artificial skin* mounted on a cylindric silicon cuff (see figure 1). The advantage of the cylindric cuff is that a scalar pressure value can be interpreted directly as a force vector perpendicular to the cylinder's surface defining the direction in which the limb should be moved.

Forces which affects the upper cuff are interpreted as forces on the elbow and forces which affects the lower cuff are interpreted as forces on the wrist. By means of these sensors, the operator can correct CORA's arm movements e.g. by lifting the elbow while the robot grasps an object, to avoid for instance an obstacle which the vision system hadn't detected or to push the endeffector out of his workspace when it disturbs the operator.

2.7 Limitations

A discussion of CORA's design would be incomplete without mentioning some major shortcomings with respect to basic human abilities. The most serious limitation is missing mobility of the robot in comparison to human. Due to the missing mobility is the restriction of the work space to table on which CORA is mounted.

3 *"Physiology"* - Hardware and System Software

The technical realization of CORA's design goals was mainly determined by the principle of using reliable "standard" hardware in order to put the emphasis on behavioral control rather than tool development.

CORA's hardware architecture meets the following main requirements:

- Realtime processing for images and arm control is possible. We determined the maximum hard realtime requirement to be about 1 ms for the arm control module. Apart from fast processors, this requires a high-bandwidth internal bus and communication system.
- The software development allows convenient access from external hosts for multiple users simultaneously. This requires a capable multiuser/multitasking OS.

The requirements were met by using a small CORA-wide network of Intel PCI PCs running QNX, a distributed realtime operating system. The network is currently realized as 100 Mbit Ethernet but will soon be changed to 1 Gbit.

3.1 Robot Arm

The arm is a modular system manufactured by Amtec. It has seven degrees of freedom, giving one degree of redundancy, which is useful for e.g. obstacle avoidance during grasping.

Each module carries a micro-controller with a CAN-bus interface. For communication with the arm an ISA-card implementing the CAN-bus protocol is used. The Arm is mounted on a one DoF trunk consisting of another *Amtec* module.

3.2 Camera Head

The camera head, consisting of an commercial wrist module and a self developed unit carrying a stereo camera system (Sony XC-999P) and microphones, provides 2 DoF (pan and tilt). Pan range is about $\pm 180^o$ and tilt movement in the range of $\pm 90^o$. The two color foveacameras with 40^o field of view are positioned with their main point at the intersection of pan and tilt axes.

3.3 The artificial skin

The artificial skin, invented by the SIEMENS robotics group, is based on a conductive foam the conductivity of which varies with the pressure applied to it. This resistance or conductivity variation is used as the output of the sensor. The variation of resistance is measured by using the

analog to digital converter of a PIC microcontroller. The sensing part of the skin is made of two layers of EVAZOTE-foam. Two electrodes are inserted in parallel on opposite sides of each layer. Those electrodes are connected to a so-called sensor board. On the upper layer one electrode is set to ground (0V) and the other to 5 volts so that the voltage varies continuously from 0 to 5 Volts within the material. The electrodes of the lower layer are both connected to the A/D converter of the microcontroller. The voltage of the lower layer is periodically varied between 0 and 5V. When force is applied to the upper layer, the voltage of the lower layer corresponds to the voltage of the upper layer depending on the position of the force so that the converter can read this position. Two analog switches are soldered on the board. They are used to set a defined calibrating voltage on the electrodes in order to measure the X and Y position. The same principle is used to measure the pressure.

3.4 Computer Network

The computational power is provided by a network of PCs - currently five 1.8GHz Pentiums - running QNX.

The master PC is central for CORA's operation: It is the first piece of hardware activated when CORA is switched on, controls all other components via an I/O card connected to various relays and acts as a gateway to the institute's LAN for external access via Ethernet. It also provides a simple status display via a couple of LEDs displaying the power and logic status of the hardware components.

4 Elementary Abilities

The main idea of CORA's behavior generation and control stems from behavior based robotics, an approach first formulated by Braitenberg 1984 [2]. The complex behavior necessary to accomplish a given task results from coupling several simple behaviors. The different simple behaviors run in parallel, each performing a meaningful action. The clever combination of these behaviors realizes complex behavior of the robot.

4.1 Adaptive Skin Color Calibration

Because of the variation of the skin color with changing lighting conditions it seemed to be necessary to compensate this by an calibration routine which determines the skin color with respect to the actual lighting conditions and the current user.

After the robot requests the operator to show his hand, the robot shots an image, transform this image into the HSI color space and masked out the backround. The backround in this case is everything except the table area. The maskerading is done by using the knowledge about the size and position of the table in the image. A region growing algorithm is used to clusters pixel with similar color. The clusters which are sufficient smaller than the expected size of the hand will be neglected. In the next step the color cluster which are not in a coarse range of the expected skin color will be also masked. The region which is nearest to the center of the image is the region which characterizes the hand. A color histogram is taken from this region to characterize the humans skin color in the color space (see Fig. 2).

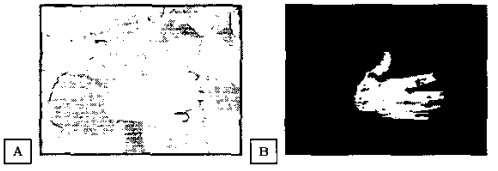

Fig. 2. Left side: The human hand during the calibration phase. Right side: Region, which represents the human skin color.

4.2 Gesture Recognition

After the initialization phase in which the operators skin color is determined (see Section 4.1), the recognition method will now be described in detail.

To locate the user's hand, we transform the image data into HSI format. Using the hue information we segment the images by using only pixels that lie in a specific small color range which has determined by the initial phase as described in Section 4.1. The bigest skin colored clusters represents the humans hand. The pixel of the bottom of this preselected hand cluster represents the fingertip of the user. This is usually for a person pointing to an object standing on the table.

Calculating the disparity, using the external and internal parameter of the cameras, for this corresponding pixel of the right and the left image, we obtain the 3-D position of the fingertip. The projection of the the fingertip position onto the table is assumed as the position the user is pointing to. The pointing direction is defined as the direction from the fingertip projection to the robot's body.

The detection of the pointing gesture serves as a preselection of the region of interest (ROI) in the image. In our case, the ROI is defined as a triangular area in the bird's eye view of the table which covers a sector with 30^0 opening angle in front of the fingertip. Within this sector the object which is nearest to the fingertip is considered

to be specified by the user's pointing gesture.

4.3 Tracking

To track the humans hand we also segment the images by using only pixels that lie in a specific small color range using the hue information as described above. For these pixels a map of binary coded edge information [4] is generated (see [5] for details). Vertical and diagonal edges in the left and right images are taken as features for stereoscopic correspondences and a greylevel based correlation measure is used to find best matches. All matches that exceed a threshold of correspondence measure are stored in a disparity map. As we use only edge information this map is sparse. To get reliable depth information regarding local objects we cluster the binarized color frames into connected regions and evaluate local disparity histograms with respect to each connected patch. For objects that do not show occlusion these disparity distributions have only one solitary peak. To manage occlusion of skin colored objects we search the peak with a minimum frequency and the highest disparity in every local patch. The peak with the highest disparity denotes the closest object. Its corresponding position in the image can be found by knowing the cluster and disparity. The object's position in the world can be estimated by triangulation if the inner and outer camera parameters are known. The complete vision process is sketched in Fig. 3.

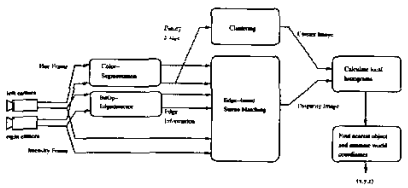

Fig. 3. Vision process

If an hand is found in the input images the head is rotated about the pan tild axis to fixate the operators hand. Since we use the stereo cameras in parallel position to fasten the vision process (by suppression of vertical disparities) to fixate means to turn the center of the cyclopean view towards the target. Image acquisition/evaluation and head rotation run asynchronously. If a new target position is calculated before the latest head-movement is finished the movement target is set to the new values. The corresponding pan/tilt-coordinates are given directly to the hardware controller which performs ramp-like velocity profiles.

4.4 Object recognition

4.4.1 Learning phase

First, the robot must be enabled to recognize a specified object. Therefore, the robot must learn some of the object's special features. In our case these features are the color distribution and the calculated bird's eye view of the object. To obtain these features, all components of the image which are not part of the table are masked. This is done by a simple geometric calculation on the basis of the known spatial relation between the camera head and the table. In a second step, the color of the table is determined by assuming the table to represent the biggest cluster in the image. Image areas with similar colors are also masked as we assume them to belong to the table's surface.

Thereby the scene gets restricted to its substantial part i.e. the object to be learned. On the basis of these limited data a color analysis and a determination of the disparity of each pixel follows. The color analysis consisting of the stored maxima of the objects color histogram, provides the characteristic color values of the object within the HSI space.

The disparity, calculated by a graylevel based algorithm, provide the data to calculate the three dimensional position of each pixel using the knowledge of the internal and external parameter of the cameras. These positions are transformed into the bird's eye view of the scene from above. This view is used as an object representation which is independent of the direction of view of the cameras and does not change its size, according to different distances. Based on this representation, the object can be found in more complex scenes later on.

4.4.2 Searching phase

In the searching phase the learned object characteristics enable a successful search for this object. Initially the scene gets reduced to its objects as done in the learning phase (i.e. masking out the table). For each cluster in the scene a color histogram is calculated. Only areas with object clusters that are sufficiently well filled with the learned color values are used in the further image processing. The calculation of the bird's eye view on the basis of the disparity information is again the same as in the learning phase.

At this time the bird's eye view of the learned object and the bird's eye view of the scene are present. In a process, similar to the Hausdorff Tracker described in [5], the bird's eye view gets mutually correlated with a blurred copy of the bird's eye views of the scene. The maximum of this correlation, verified by an inverse correlation, specifies the most probable position of the object

in the scene. Because the object can be rotated in the two dimensional plane, the correlation must be accomplished for a number of possible rotation angles. In the application on CORA rotation steps of 15^o, depending on the object size, turned out to be sufficient. If the corelation of the searched object is lower than the given threshold the system notifies that the scene doesn't contain the searched object and in addition to it the system also notifies if the table is empty or not.

The most important advantaged of this algorithm are that the learned object even can be find if it isn't well separated from the other objects in the scene (i.e. it touches an other object) and that also the infomation about its position and orientation is provided.

Fig. 4. Figure (a) shows a typical situation during human robot interaction. The segmentation and clustering of the human skin can be seen in figure (b). The right figure (c) shows the disparity image.

4.5 Speech Recognition

Using the speech recognition system *ears* [7] [8] spoken keyword commands can be identified . By means of the speech synthesizer *mbrola* [1], the robot can express its behavioral state. On the basis of natural commands it is possible to guide the robot's manipulator to a certain target or to terminate an incorrect behavior (such as the selection of a wrong object) by a spoken command.

4.6 Grasping

The flexible grasping in the whole work space is realized by implementing a dynamical manipulator control. The implemented control scheme includes the solution of the inverse kinematics problem of an 8 DoF manipulator [6] and the dynamic control of the endeffector under obstacle avoidance [6].

5 Experimental Results

To demonstrate the functionality of the basic behaviors described above the organization of the several elementary behaviors, is done algorithmic. It is an ongoing work to generate and organize the behaviors in this scenario using the so called *Dynamic Approach to Robotics* [9] based on dynamical systems.

By means of an assembly task, we show the integration of the implemented interaction channels, described above. The human user should assembly together with the robot assistant a work piece consisting of two parts. After the human user starts with the spoken command *init* the session, the robot initializes his hardware (framegrabber, cameras, microphones, arm etc.) and gives a spoken report about his internal state.

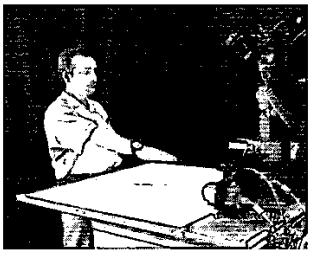

Fig. 5. Skin color calibration

After that the robot requests the user by speech to show his hand to calibrate the skin color (see Fig. 5). The robot assistant confirms the skin color calibration by speech and tells the human user, that he is now ready to assist him.

Fig. 6. The robot starts to grasp the desired object and the human user correct the grasping orientation by touching and turning the gripper

At this point the robot, has just the ability to recognize speech, to detect a pointing gesture, to track the humans hand and to grasp generic unknown objects from the table. To allow the robot assistant to recognize objects, the human user initiates an learning phase by the keyword command *learn* and put then the object to be learned on the table.

Fig. 7. Overhand the object

The robot shots an image of the object, extracts and stores the features as described in Section 4.4. Starting with an arbitrary scene consisting of several objects and tools, the human user will now try to assembly the work piece. First the human gives the command *grasp* without any additional hint. The robot searches the scene and

find in this case just one known object (i.e one of the objects he has learned before), which means that no ambiguities occurs in his internal representation and therefor he starts grasping the learned object (see Fig. 6).

Afterwards the robot search and track the humans hand to overhand the object (see Fig. 7). Evaluating the force/torque sensor the robot detect the contact with the humans hand and opens the gripper. To assembly the second part of the

Fig. 8. Pointing

work piece with the one the robot has overhanded the human user, a tool is needed. The user just gives again the command *grasp*. The robot in starts again searching the scene but this time he can't find any known object. At this point the robot requests the user to disambiguates the situation by pointing on the desired object. While the user is pointing on the tool the robot starts first tracking the users hand and when the hand stops moving the robot extracts and confirms by speech the detection of the pointing gesture (see Fig. 8). Again the robot grasp the object and

Fig. 9. Grasping and over handing the tool

overhanded it to the human user, who assembly now the to parts of the work piece (see Fig. 9).

At the end of the session the user speaks the command *clean up* and the robot put all objects from the table in the box besides him. After clea-

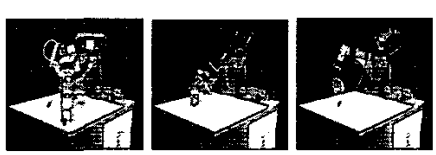

Fig. 10. Cleaning up the table

ning up the table the robot notifies by speech the end of the session.

6 Conclusion and outlook

This paper describes the design of CORA, an autonomous robot assistant for service tasks. Since service robots have to perform well in a common environment with humans we propose that these robots have to be designed as anthropomorphic as possible.

The best way to achieve complex behavior consists in coupling of simple behaviors. This will be done best by designing each behavior as a dynamical system and coupling these system by sensor data.

In the future we implement the whole range of behaviors necessary to cope with the discussed scenarios.

The discussion of CORA's behavioral architecture, i.e. our methods to combine relatively simple behavioral modules into complex systems, is merely sketched in this paper and will be addressed in future publications.

Acknowledgment

This work is supported by the BMBF grant MORPHA (925 52 12).

References

[1] B Bozkurt, M Bagein, and T Dutoit. From mbrola to nu-mbrola. In *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis Blair Atholl, Scotland, 2001*, pages 127–130, 2001.

[2] V. Braitenberg. *Vehicles. Experiments in Synthetic Psychology*. MIT Press, Cambridge, Mass., 1984.

[3] Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer, 1997.

[4] Christian Goerick. Local orientation coding and adaptive thresholding for real time early vision. Technical report, Institut für Neuroinformatik, Lehrstuhl für Theoretische Biologie, Ruhr-Universität Bochum, 1994.

[5] D.P. Huttenlocher, G.A. Klanderman, and W.J.Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. on PAMI*, PAMI-15-9:850–863, 1993.

[6] I. Iossifidis and A. Steinhage. Control of an 8 dof manipulator by means of neural fields. In *FSR2001, International Conference on Field and Service Robotics*, Helsinki, Finland, 2001.

[7] Defense Advanced Research Projects Agency: Information Processing Technology Office (IPTO). Planned procurements, December 2001.

[8] S E Johnson, P Jourlin, G L Moore, K Sprck Jones, and P C Woodland. The cambridge university spoken document retrieval system. In *Proc ICASSP '99*, volume 1, pages 49–52, Phoenix, AZ, 1999.

[9] G. Schöner, M. Dose, and C. Engels. Dynamics of behavior: Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16:213–245, 1995.